Package ‘PRIMsrc’

October 6, 2021

Type Package

Title PRIM Survival Regression Classification

Version 0.9.0

Date 2021-10-06

Author Jean-Eudes Dazard [aut, cre], Michael Choe [ctb], Michael LeBlanc [ctb], Alberto Santana [ctb], J. Sunil Rao [ctb]

Maintainer Jean-Eudes Dazard <jean-eudes.dazard@case.edu>

Description Performs a unified treatment of Bump Hunting by Patient Rule Induction Method (PRIM) in Survival, Regression and Classification settings (SRC) in a multivariate settings and in high-dimensional data. The current version is a development release that only implements the case of a survival response.

Depends R (>= 3.5.0)

Imports parallel, survival, glmnet, superpc, Hmisc, quantreg

NeedsCompilation yes

URL https://github.com/jedazard/PRIMsrc, https://www.primsrc.com

Repository CRAN, GitHub, Inc.

Date/Publication 2015-07-28

License GPL (>= 3) | file LICENSE

Archs i386, x64

R topics documented:

`PRIMsrc-package` *Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification in a multivariate setting and in high-dimensional data*

Description

"Bump Hunting" (BH) refers to the procedure of mapping out a local region of the multi-dimensional input space where a target function of interest, usually unknown, assumes smaller or larger values than its average over the entire space. In general, the region R could be any smooth shape (e.g. a convex hull) possibly disjoint.

PRIMsrc implements a unified treatment of "Bump Hunting" (BH) by algorithms derived from the Patient Rule Induction Method (PRIM) (Friedman and Fisher, 1999) for Survival, Regression and Classification outcomes (SRC). To estimate the region, PRIM generates decision rules delineating hyperdimensional boxes (hyperrectangles) of the input space, not necessarily contiguous, where the outcome is smaller or larger than its average over the entire space.

Assumptions are that the multivariate input variables can be discrete or continuous, and the univariate outcome variable can be discrete (Classification), continuous (Regression), or a time-to-event, possibly censored (Survival). It is intended to handle low and high-dimensional multivariate datasets, including the paradigm where the number of covariates p exceeds or dominates that of samples ($p > n$ or $p \gg n$).

Please note that the current version is a development release, that only implements the case of a survival outcome. At this point, this version of ‘PRIMsrc‘ is also restricted to a directed peeling search of the first box covered by the recursive coverage (outer) loop of our PRSP or PRGSP algorithm (see details below). New features will be added as soon as available.

Details

In a direct application, "Bump Hunting" (BH) can identify subgroups of observations for which their outcome is as extreme as possible. Similarly to this traditional goal of subgroup finding, **PRIMsrc** also implements the alternative goal of mapping out a region (possibly disjointed) of the input space where the outcome *difference* between existing (fixed) groups of observations is as extreme as possible. We refer to the later goal as "Group Bump Hunting" (GBH).

In the case of a time-to event outcome, possibly censored (as in survival or risk analysis), "Survival Bump Hunting" (SBH) is done by our Patient Recursive Survival Peeling (PRSP) algorithm. See Dazard and Rao (2014, 2015, 2016, 2021a) for details, as well as Dazard et al. (2021c) for an application in Patient Survival Subtyping. Alternatively, "Group Survival Bump Hunting" (GSBH) is done by using a derivation of PRSP with specifc peeling and cross-validation criterion, called Patient Recursive Group Survival Peeling (PRGSP). See Dazard and Rao (2021b) for details, as well as Rao et al. (2021d) for an application in Survival Disparity Subtyping.

The package relies on an optional variable screening (pre-selection) procedure that is run before the PRSP algorithm and final variable usage (selection) procedure is done. This is done by four possible cross-validated variable screening (pre-selection) procedures offered to the user from the main end-user survival Bump Hunting function `sbh`. See Dazard and Rao (2014, 2015, 2016, 2021a, 2021b) for details, as well as Dazard et al. (2021c) and Rao et al. (2020) for applications in Patient Survival Subtyping and Disparity Subtyping.

The following describes the end-user functions that are needed to run a complete procedure. The other internal subroutines are not documented in the manual and are not to be called by the end-user at any time. For computational efficiency, some end-user functions offer a parallelization option that is done by passing a few parameters needed to configure a cluster. This is indicated by an asterisk (* = optionally involving cluster usage). The R features are categorized as follows:

1. END-USER SURVIVAL BUMP HUNTING FUNCTION
   `sbh` * **Cross-Validated Survival Bump Hunting**
   Main end-user function for fitting a Survival Bump Hunting (SBH) model. It returns an object of class `sbh`, as generated by our Patient Recursive Survival Peeling (PRSP) algorithm (or Patient Recursive Group Survival Peeling (PRGSP)), containing cross-validated estimates of the target region (bump) of the input space with end-points statistics of interest. The main function relies on an optional internal variable screening (pre-selection) procedure that is run before the actual variable usage (selection) is done at the time of fitting the Survival Bump Hunting (SBH) or Group Survival Bump Hunting (GSBH) model model using our PRSP or PRGSP algorithm. At this point, the user can choose between four possible variable screening (pre-selection) procedures:

   (a) Univariate Patient Recursive Survival Peeling algorithm (default of package PRIMsrc)
   (b) Penalized Censored Quantile Regression (by Semismooth Newton Coordinate Descent algorithm adapted from package hqreg)
   (c) Penalized Partial Likelihood (adapted from package glmnet)
   (d) Supervised Principal Component Analysis (adapted from package superpc)

   In this version, the Cross-Validation (CV) that controls model size (#covariates) and model complexity (#peeling steps), respectively, to fit the Survival Bump Hunting model, are carried out internally by two consecutive tasks within the single main function `sbh()`. The returned S3-class `sbh` object contains cross-validated estimates of all the decision-rules of used covariates and all other statistical quantities of interest at each iteration of the peeling sequence (inner loop of the PRSP algorithm). This enables the graphical display of results of profiling curves for model selection/tuning, peeling trajectories, covariate traces and survival distributions (see plotting functions for more details). The function offers a number of options for the number of replications of the fitting procedure to be perfomed: B; the type of K-fold cross-validation desired: (replicated)-averaged or-combined; as well as the peeling and cross-validation critera for model selection/tuning, and a few more parameters for the PRSP algorithm. The function takes advantage of the R packages **parallel** and **snow**, which allows users to create a parallel backend within an R session, enabling access to a cluster of compute cores and/or nodes on a local and/or remote machine(s) with either. **PRIMsrc** supports two types of communication mechanisms between master and worker processes: 'Socket' or 'Message-Passing Interface' ('MPI').

2. END-USER FUNCTION FOR CONTROLLING ANCILLARY MODEL PARAMETERS
   `sbh.control` **Parameters Control Function**
   End-user function to set ancillary parameters of main end-user function `sbh` for fitting a Survival Bump Hunting (SBH) model.

3. END-USER FUNCTION FOR PACKAGE NEWS
   `PRIMsrc.news` **Display the PRIMsrc Package News**

End-user function to display the log file `NEWS` of updates of the **PRIMsrc** package.

4. END-USER S3-METHOD FUNCTIONS FOR SUMMARY, DISPLAY, PLOT AND PREDICTION

   `summary` **Summary Function**
   End-user S3-method summary function to summarize the main parameters used to generate the `sbh` object.

   `print` **Print Function**
   End-user S3-method print function to display the cross-validated estimated values of the `sbh` object.

   `plot` **2D Visualization of Data Scatter and Encapsulating Box**
   End-user S3-method plot function for two-dimensional visualization of scatter of data points and cross-validated encapsulating box of a `sbh` object for the higher-risk (in-bump) versus lower-risk (out-bump) bumps (bump difference, PRSP algorithm), or between two specified fixed groups (group difference, PRGSP algorithm), if this option is used. The scatter plot is done for a given peeling step (or number of steps) of the peeling sequence (inner loop of our PRSP or PRGSP) and in a given plane of the used covariates of the `sbh` object, both specified by the user.

   `predict` **Predict Function**
   End-user S3-method predict function to predict the box membership and box vertices on an independent set.

5. END-USER PLOTTING FUNCTIONS FOR MODEL VALIDATION AND VISUALIZATION OF RESULTS

   `plot_profile` **Visualization for Model Selection/Validation**
   End-user function for plotting the cross-validated model selection/tuning profiles of a `sbh` object. It uses the user's choice of cross-validation criterion statistics among the Log Hazard Ratio (LHR), Log-Rank Test (LRT) or Concordance Error Rate (CER). The function plots (as it applies) both profiles of cross-validation criterion as a function of variables screening size (cardinal subset of top-screened variables in the PRSP variable screening procedure), and peeling length (number of peeling steps of the peeling sequence in the inner loop of the PRSP or PRGSP algorithm).
   `plot_traj` **Visualization of Peeling Trajectories/Profiles**
   End-user function for plotting the cross-validated peeling trajectories/profiles of a `sbh` object. Applies to the user-specified covariates among the pre-selected ones and all other statistical quantities of interest at each iteration of the peeling sequence (inner loop of our PRSP or PRGSP algorithm).
   `plot_trace` **Visualization of Covariates Traces**
   End-user function for plotting the cross-validated covariates traces of a `sbh` object. Plot the cross-validated modal trace curves of covariate importance and covariate usage of the user-specified covariates among the pre-selected ones at each iteration of the peeling sequence (inner loop of our PRSP or PRGSP algorithm).
   `plot_km` **Visualization of Survival Distributions**
   End-user function for plotting the cross-validated survival distributions of a `sbh` object. It plots the cross-validated Kaplan-Meir estimates of survival distributions either between higher-risk (in-bump) versus lower-risk (out-bump) bumps of observations (bump difference, PRSP algorithm), or between two specified fixed groups (group difference, PRGSP algorithm). The plot is done for a given peeling step (or number of steps) of the peeling sequence (inner loop of

our PRSP or PRGSP) algorithm) of the sbh object.

6. END-USER DATASETS
   Synthetic.1, Synthetic.1b, Synthetic.2, Synthetic.3, Synthetic.4 **Five Datasets From Simulated Regression Survival Models**
   Five datasets from simulated regression survival models #1-4 as described in Dazard et al. (2015), representing low- and high-dimensional situations, and where regression parameters represent various types of relationship between survival times and covariates including saturated and noisy situations. In three datasets where non-informative noisy covariates were used, these covariates were not part of the design matrix (models #2-3 and #4). In one dataset, the signal is limited to a box-shaped region R of the predictor space (model #1b). In the last dataset, the signal is limited to 10% of the predictors in a $p > n$ situation (model #4). See each dataset for more details.

   Real.1 **Clinical Dataset**
   Publicly available HIV clinical data from the Women's Interagency HIV cohort Study (WIHS). The entire study enrolled 1164 women. Inclusion criteria of the study are: women at enrolment must be (i) alive, (ii) HIV-1 infected, and (iii) free of clinical AIDS symptoms. Women were followed until the first of the following occurred: (i) treatment initiation (HAART), (ii) AIDS diagnosis, (iii) death, or administrative censoring. The studied outcomes were the competing risks "AIDS/Death (before HAART)" and "Treatment Initiation (HAART)". However, for simplification purposes, only the first of the two competing events (i.e. the time to AIDS/Death), was used. Likewise, for simplification in this clinical dataset example, only $n = 485$ complete cases were used. Variables included history of Injection Drug Use ("IDU") at enrollment, African American ethnicity ('Race'), age ('Age'), and baseline CD4 count ('CD4'). The question in this dataset example was whether it is possible to achieve a prognostication of patients for AIDS and HAART. See dataset documentation for more details.

   Real.2 **Genomic Dataset**
   Publicly available lung cancer genomic data from the Chemores Cohort Study. This data is part of an integrated study of mRNA, miRNA and clinical variables to characterize the molecular distinctions between squamous cell carcinoma (SCC) and adenocarcinoma (AC) in Non Small Cell Lung Cancer (NSCLC) aside large cell lung carcinoma (LCC). Tissue samples were analysed from a cohort of 123 patients, who underwent complete surgical resection at the Institut Mutualiste Montsouris (Paris, France) between 30 January 2002 and 26 June 2006. The studied outcome was the "Disease-Free Survival Time". Patients were followed until the first relapse occurred or administrative censoring. In this genomic dataset, the expression levels of Agilent miRNA probes ($p = 939$) were included from the $n = 123$ cohort samples. In addition to the genomic data, five clinical variables, also evaluated on the cohort samples, are included as continuous variable ('Age') and nominal variables ('Type','KRAS.status','EGFR.status','P53.status'). This dataset represents a situation where the number of covariates dominates the number of complete observations, or $p >> n$ case. See dataset documentation for more details.

Known Bugs/Problems : None reported at this time.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.
- Bacon M.C., von Wyl V., Alden C. et al. (2005). "*The Women's Interagency HIV Study: an observational cohort brings clinical sciences to the bench.*" Clin. Diagn. Lab. Immunol., 12(9):1013-1019.
- Lazar V. et al. (2013). "*Integrated molecular portrait of non-small cell lung cancers.*" BMC Medical Genomics 6:53-65.

See Also

- R package **parallel**
- R package **glmnet**
- R package **hqreg**
- R package **superpc**

plot.sbh *2D Visualization of Data Scatter and Encapsulating Box*

Description

S3-method plot function for two-dimensional visualization of scatter of data points and cross-validated encapsulating box of a sbh object. The box represents a Bump Hunting search either between higher-risk (in-bump) versus lower-risk (out-bump) observations (bump difference, PRSP algorithm), or between two specified fixed groups (group difference, PRGSP algorithm). The scatter plot is done for a given peeling step (or number of steps) of the peeling sequence (inner loop of our PRSP or PRGSP), and in a given projection plane of the used covariates of the sbh object, both specified by the user.

Usage

```
  ## S3 method for class 'sbh'
plot(x,
     main = "Scatter Plot",
     proj = x$cvfit$cv.used[c(1,2)],
     steps = x$cvfit$cv.nsteps,
     pch = 16,
     cex = 0.5,
     col = c(1,2),
     boxes = TRUE,
     asp = NA,
     col.box = rep(2,length(steps)),
     lty.box = rep(2,length(steps)),
     lwd.box = rep(1,length(steps)),
     add.caption.box = boxes,
     text.caption.box = paste("Step: ", steps, sep=""),
     pch.group = c(1,1),
     cex.group = c(1,1),
     col.group = c(3,4),
     add.caption.group = ifelse(test = (x$cvarg$peelcriterion == "grp"),
                                yes = TRUE,
                                no = FALSE),
     text.caption.group = levels(x$groups),
     device = NULL,
     file = "Scatter Plot",
     path = getwd(),
     horizontal = FALSE,
     width = 5,
     height = 5, ...)
```

Arguments

x Object of class sbh as generated by the main function sbh.

main Character vector. Main Title. Defaults to "Scatter Plot".

proj Integer vector of length two, specifying the two used (selected) covariates in which the scatter plot is to be plotted. See details. Defaults to first two used (selected) covariates.

`steps` `Integer vector`. Vector of peeling steps at which to plot the inbox samples and box vertices. Defaults to the last peeling step of `sbh` object `x`.

`pch` `Integer` scalar specifying the symbol for the outbox and inbox data points (Defaults to 16 for both).

`cex` `Numeric` scalar specifying the symbol expansion for the outbox and inbox data points (Defaults to 0.5 for both).

`col` `Integer vector` specifying the symbol color for the outbox and inbox data points (Defaults to "black" and "red", respectively).

`boxes` `Logical` scalar. Shall the encapsulating box(es) be plotted as well? Default to `TRUE`.

`asp` `Numeric` scalar giving the y/x aspect ratio. Default to `asp=NA` i.e. a regular plot. See details.

`col.box` `Integer vector` of line color of box vertices for each step. Defaults to vector of 2's (red) of length the number of steps. The vector is reused cyclically if it is shorter than the number of steps.

`lty.box` `Integer vector` of line type of box vertices for each step. Defaults to vector of 2's of length the number of steps. The vector is reused cyclically if it is shorter than the number of steps.

`lwd.box` `Integer vector` of line width of box vertices for each step. Defaults to vector of 1's of length the number of steps. The vector is reused cyclically if it is shorter than the number of steps.

`add.caption.box`
`Logical` scalar. Shall the caption be plotted? Defaults to `boxes` value.

`text.caption.box`
`Character vector` of caption content. Defaults to `paste("Step: ",steps,sep="")`.

`pch.group` `Integer vector` specifying the symbol for the two groups data points (Defaults to 0.5 for both).

`cex.group` `Numeric vector` specifying the symbol expansion for the two groups data points (Defaults to 0.5 for both).

`col.group` `Integer vector` specifying the symbol color for the two groups data points (Defaults to "green" and "blue").

`add.caption.group`
`Logical` scalar. Shall the caption be plotted? Defaults to `TRUE` or `FALSE`, depending on `x$cvarg$peelcriterion`.

`text.caption.group`
`Character vector` of caption content. Defaults to `levels(x$groups)`.

`device` Graphic display device in {`NULL`, "PS", "PDF"}. Defaults to `NULL` (standard output screen). Currently implemented graphic display devices are "PS" (Postscript) or "PDF" (Portable Document Format).

`file` File name for output graphic. Defaults to "Scatter Plot".

`path` Absolute path (without final (back)slash separator). Defaults to working directory path.

`horizontal` `Logical` scalar. Orientation of the printed image. Defaults to `FALSE`, that is potrait orientation.

`width` `Numeric` scalar. Width of the graphics region in inches. Defaults to 5.

`height` `Numeric` scalar. Height of the graphics region in inches. Defaults to 5.

`...` Generic arguments passed to other plotting functions.

Details

Use graphical parameter `asp=1` for a plotting a proportional scatter plot on the graphical device with geometrically equal scales on the x and y axes. In that case, it produces a proportional scatter plot where distances between points are represented accurately on screen. The window is set up so that one data unit in the x direction is equal in length to one data unit in the y direction.

The two dimensions (`proj`) of the projection plane in which the scatter plot is to be plotted, must be a subset (in the large sense) of the used (selected) covariates of `sbh` object `x`. If the number of used covariates in the `sbh` object is zero, the scatterplot will not be plotted. If the number of used covariates is one, the scatterplot will be plotted using the specified covariate and an arbitrary dimension, both specified by the user.

Value

Invisible. None. Displays the plot(s) on the specified `device`.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user plotting function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.

- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

plot_km *Visualization of Survival Distributions*

Description

Function for plotting the cross-validated survival distributions of a sbh object. It plots the cross-validated Kaplan-Meir estimates of survival distributions either between higher-risk (in-bump) versus lower-risk (out-bump) bumps of observations (bump difference, PRSP algorithm), or between two specified fixed groups (group difference, PRGSP algorithm). The plot is done for a user-specified number of steps of the peeling sequence, i.e. peeling step of the inner loop of the Patient Recursive Survival Peeling (PRSP) or of the Patient Recursive Group Survival Peeling (PRGSP) algorithm of the sbh object.

Usage

```
plot_km(object,
        main = "Survival KM Plots",
        xlab = "Time",
        ylab = "Probability",
        ci = TRUE,
        precision = 1e-3,
        mark = 3,
        col = c(1,2),
        lty = 1,
        lwd = 0.5,
        cex = 0.5,
        steps = 1:object$cvfit$cv.nsteps,
        plot.type = "bumps",
        bump.reference = "in-bump",
        group.reference = levels(object$groups)[1],
        add.caption = TRUE,
        text.caption = c("out-bump","in-bump"),
        nr = 3,
        nc = 4,
        device = NULL,
        file = "Survival KM Plots",
        path = getwd(),
        horizontal = TRUE,
        width = 11,
        height = 8.5, ...)
```

Arguments

`object` Object of class `sbh` as generated by the main function `sbh`.

`main` `Character vector`. Main Title. Defaults to "Survival KM Plots".

`xlab` `Character vector`. X-axis label. Defaults to "Time".

`ylab` `Character vector`. Y-axis label. Defaults to "Probability".

`ci` `Logical` scalar. Shall the 95% confidence interval be plotted? Defaults to `TRUE`.

`precision` Precision of log-rank p-values of separation between two survival curves. Defaults to 1e-3.

`mark` `Integer` scalar of mark parameter, which will be used to label the inbox and outbox curves. Defaults to 3.

`col` `Integer` scalar specifying the color of the inbox and outbox curves (Defaults to `c(1,2)`).

`lty` `Integer` scalar. Line type for the survival curve. Defaults to 1.

`lwd` `Numeric` scalar. Line width for the survival curve. Defaults to 0.5.

`cex` `Numeric` scalar specifying the size of the marks, symbol expansion used for titles, captions, and axis labels. Defaults to 0.5.

`steps` `Integer vector`. Vector of peeling steps at which to plot the survival curves. Defaults to all the peeling steps of `sbh` object `object`.

`plot.type` `Character vector` of plot type in {"bumps", "groups"} for plotting survival curves between ("out-bump","in-bump") or between the fixed groups (`levels`(object$groups)), respectively. Defaults to "bumps".

`bump.reference` `Character vector` in {"out-bump","in-bump"} of which bump is taken as the reference for plotting survival curves. Defaults to "in-bump".

`group.reference`

`Character vector` of which of the fixed groups is taken as the reference for plotting survival curves. Defaults to `levels`(object$groups)[1].

`add.caption` `Logical` scalar. Shall the caption be plotted? Defaults to `TRUE`.

`text.caption` `Character vector` of caption content. Defaults to {"out-bump","in-bump"}.

`nr` `Integer` scalar of the number of rows in the plot. Defaults to 3.

`nc` `Integer` scalar of the number of columns in the plot. Defaults to 4.

`device` `Character vector` of graphic display device in {`NULL`, "PS", "PDF"}. Defaults to `NULL` (standard output screen). Currently implemented graphic display devices are "PS" (Postscript) or "PDF" (Portable Document Format).

`file` `Character vector` of file name for output graphic. Defaults to "Survival Plots".

`path` `Character vector` of absolute path (without final (back)slash separator). Defaults to the working directory path.

`horizontal` `Logical` scalar. Orientation of the printed image. Defaults to `TRUE`, that is potrait orientation.

`width` `Numeric` scalar. Width of the graphics region in inches. Defaults to 11.

`height` `Numeric` scalar. Height of the graphics region in inches. Defaults to 8.5.

`...` Generic arguments passed to other plotting functions, including `plot.survfit` (R package **survival**).

Details

Some of the plotting parameters are further defined in the function `plot.survfit` (R package **survival**). Step #0 always corresponds to the situation where the starting box covers the entire test-set data before peeling.

The plot is done for the given peeling criterion (`object$cvarg$peelcriterion`) of the `sbh` object. If a regular hunt of *bump difference* is done (`peelcriterion` in {"lrt", "lhr", "chs"}), cross-validated Kaplan-Meir estimates (KM curves) are plotted between observations from the highest risk bump (in-bump) versus lower-risk bump (out-bump). If a hunt of (user-specified) fixed *group difference* is done (`peelcriterion` in {"bwgrp", "bwbmp"}), KM curves are plotted either: (i) between observations of both groups within the highest risk bump (in-bump) ("bwgrp"), or similarly, (ii) between observations from the highest risk bump (in-bump) versus lower-risk bump (out-bump) within a given group ("bwbmp").

Cross-validated LRT, LHR values and log-rank p-values of separation between bumps or groups are shown at the bottom of the plot with the corresponding peeling step. P-values are lower-bounded by the precision limit given by $1/A$, where A is the number of permutations.

Value

Invisible. None. Displays the plot(s) on the specified `device`.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user plotting function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <`jean-eudes.dazard@case.edu`>
- "Michael Choe, M.D." <`mjc206@case.edu`>
- "Michael LeBlanc, Ph.D." <`mleblanc@fhcrc.org`>
- "Alberto Santana, MBA." <`ahs4@case.edu`>
- "J. Sunil Rao, Ph.D." <`Rao@biostat.med.miami.edu`>

Maintainer: "Jean-Eudes Dazard, Ph.D." <`jean-eudes.dazard@case.edu`>

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.

- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression*." J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

See Also

- `plot.survfit` (R package **survival**)

`plot_profile` *Visualization for Model Selection/Tuning*

Description

Function for plotting the cross-validated model selection/tuning profiles of a sbh object. It uses the user's choice of cross-validation criterion statistics among the Log Hazard Ratio (LHR), Log-Rank Test (LRT) or Concordance Error Rate (CER). The function plots (as it applies) both profiles of cross-validation criterion as a function of variables screening size (cardinal subset of top-screened variables in the PRSP variable screening procedure), and peeling length (number of peeling steps of the peeling sequence in the inner loop of the PRSP or PRGSP algorithm).

Usage

```
plot_profile(object,
             main = "Profile Plots",
             xlim = NULL,
             ylim = NULL,
             add.sd = TRUE,
             add.profiles = TRUE,
             add.caption = TRUE,
             text.caption = c("Mean","Std. Error"),
             pch = 20,
             col = 1,
             lty = 1,
             lwd = 0.5,
             cex = 0.5,
             device = NULL,
             file = "Profile Plots",
             path = getwd(),
             horizontal = FALSE,
             width = 8.5,
             height = 5.0, ...)
```

Arguments

`object` Object of class `sbh` as generated by the main function `sbh`.

`main` `Character vector`. Main Title. Defaults to "Profile Plots".

`xlim` `Numeric vector` of length 2. The x limits [x1, x2] of the plot. Defaults to `NULL`.

`ylim` `Numeric vector` of length 2. The y limits [y1, y2] of the plot. Defaults to `NULL`.

`add.sd` `Logical` scalar. Shall the standard error bars be plotted? Defaults to `TRUE`.

`add.profiles` `Logical` scalar. Shall the individual profiles (for all replicates) be plotted? Defaults to `TRUE`.

`add.caption` `Logical` scalar. Should the caption be plotted? Defaults to `TRUE`.

`text.caption` `Character vector` of caption content. Defaults to {"Mean","Std. Error"}.

`pch` `Integer` scalar of symbol number for all the profiles. Defaults to 20.

`col` `Integer` scalar of line color of the mean profile. Defaults to 1.

`lty` `Integer` scalar of line type of the mean profile. Defaults to 1.

`lwd` `Numeric` scalar of line width of the mean profile. Defaults to 0.5.

`cex` `Numeric` scalar of symbol expansion for all the profiles. Defaults to 0.5.

`device` Graphic display device in {`NULL`, "PS", "PDF"}. Defaults to `NULL` (standard output screen). Currently implemented graphic display devices are "PS" (Postscript) or "PDF" (Portable Document Format).

`file` File name for output graphic. Defaults to "Profile Plot".

`path` Absolute path (without final (back)slash separator). Defaults to working directory path.

`horizontal` `Logical` scalar. Orientation of the printed image. Defaults to `FALSE`, that is potrait orientation.

`width` `Numeric` scalar. Width of the graphics region in inches. Defaults to 8.5.

`height` `Numeric` scalar. Height of the graphics region in inches. Defaults to 5.0.

`...` Generic arguments passed to other plotting functions.

Details

Model tuning is done by applying the cross-validation criterion defined by the user's choice of specific statistic. The goal is to find the optimal value of model parameters by maximization of LHR or LRT, or minimization of CER. The parameters to optimize are (i) the cardinal of top-ranked variables subsets (if the "prsp" variable screening is chosen), and (ii) the number of peeling steps of the peeling sequence (inner loop of our PRSP algorithm) in any case of variable screening method.

Currently, this is done internally for visualization purposes, but it will ultimately offer the option to be done interactively with the end-user as well for parameter choosing/model selection.

Value

Invisible. None. Displays the plot(s) on the specified `device`.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user plotting function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

---

plot_trace *Visualization of Covariates Traces*

Description

Function for plotting the cross-validated covariates traces of a sbh object. Plot the cross-validated modal trace curves of covariate importance and covariate usage of the pre-selected covariates specified by user at each iteration of the peeling sequence (inner loop of our PRSP or PRGSP algorithm).

Usage

```
plot_trace(object,
           main = "Covariate Trace Plots",
           xlab = expression(paste("Box Support (", beta, ")", sep="")),
           ylab = "Covariate Range (centered)",
           toplot = object$cvfit$cv.used,
           center = TRUE,
           scale = FALSE,
           col.cov,
           lty.cov,
           lwd.cov,
           col = 1,
           lty = 1,
           lwd = 0.5,
           cex = 0.5,
           add.caption = FALSE,
           text.caption = NULL,
           device = NULL,
           file = "Covariate Trace Plots",
           path = getwd(),
           horizontal = FALSE,
           width = 8.5,
           height = 8.5, ...)
```

Arguments

`object` Object of class `sbh` as generated by the main function `sbh`.

`main` `Character vector`. Main Title. Defaults to "Covariate Trace Plots".

`xlab` `Character vector`. X-axis label. Defaults to "Box Support (beta)". `NULL`

`ylab` `Character vector`. Y-axis label. Defaults to "Covariate Range (centered)".

`toplot` `Numeric vector`. Which of the pre-selected covariates to plot (in reference to the original index of covariates). Defaults to covariates used for peeling.

`center` `Logical` scalar. Shall the data be centered?. Defaults to `TRUE`.

`scale` `Logical` scalar. Shall the data be scaled? Defaults to `FALSE`.

`col.cov` `Integer vector`. Line color for the covariate importance curve of each selected covariate. Defaults to vector of colors of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`lty.cov` `Integer vector`. Line type for the covariate importance curve of each selected covariate. Defaults to vector of 1's of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`lwd.cov` `Integer vector`. Line width for the covariate importance curve of each selected covariate. Defaults to vector of 1's of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`col` `Integer` scalar. Line color for the covariate trace curve. Defaults to 1.

`lty` `Integer` scalar. Line type for the covariate trace curve. Defaults to 1.

`lwd` `Numeric` scalar. Line width for the covariate trace curve. Defaults to 0.5.

`cex` `Numeric` scalar. Symbol expansion used for titles, captions, and axis labels. Defaults to 0.5.

`add.caption` `Logical` scalar. Should the caption be plotted?. Defaults to `FALSE`.

`text.caption` `Character vector` of caption content. Defaults to `NULL`.

`device` Graphic display device in {`NULL`, "PS", "PDF"}. Defaults to `NULL` (standard output screen). Currently implemented graphic display devices are "PS" (Postscript) or "PDF" (Portable Document Format).

`file` File name for output graphic. Defaults to "Covariate Trace Plots".

`path` Absolute path (without final (back)slash separator). Defaults to working directory path.

`horizontal` `Logical` scalar. Orientation of the printed image. Defaults to `FALSE`, that is potrait orientation.

`width` `Numeric` scalar. Width of the graphics region in inches. Defaults to 8.5.

`height` `Numeric` scalar. Height of the graphics region in inches. Defaults to 8.5.

`...` Generic arguments passed to other plotting functions.

Details

The trace plots limit the display of traces to those only covariates that are used for peeling. If centered, an horizontal black dotted line about 0 is added to the plot.

Due to the variability induced by cross-validation and replication, it is possible that more than one covariate be used for peeling at a given step. So, for simplicity of the trace plots, only the modal or majority vote trace value (over the folds and replications of the cross-validation) is plotted.

The top plot shows the overlay of covariate importance curves for each covariate. The bottom plot shows the overlay of covariate usage curves for each covariate. It is a dicretized view of covariate importance.

Both point to the magnitude and order with which covariates are used along the peeling sequence.

Value

Invisible. None. Displays the plot(s) on the specified `device`.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user plotting function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

plot_traj *Visualization of Peeling Trajectories/Profiles*

Description

Function for plotting the cross-validated peeling trajectories/profiles of a sbh object. Applies to the pre-selected covariates specified by user and other output statistics of interest at each iteration of the peeling sequence (inner loop of our PRSP or PRGSP algorithm).

Usage

```
plot_traj(object,
          main = "Trajectory Plots",
          toplot = object$cvfit$cv.used,
          range = NULL,
          col.cov,
          lty.cov,
          lwd.cov,
          col = 1,
          lty = 1,
          lwd = 0.5,
          cex = 0.5,
          add.caption = FALSE,
          text.caption = NULL,
          nr = NULL,
          nc = NULL,
          device = NULL,
          file = "Trajectory Plots",
          path = getwd(),
          horizontal = FALSE,
          width = 8.5,
          height = 11, ...)
```

Arguments

`object` Object of class `sbh` as generated by the main function `sbh`.

`main` `Character vector`. Main Title. Defaults to "Trajectory Plots".

`toplot` `Numeric vector`. Pre-selected covariates to plot (in reference to the original index of covariates). Defaults to covariates used for peeling.

`range` `List`. User-specified ranges of survival output statistics according to the peeling criterion (`peelcriterion`) that is used, so that only relevant entries are to be entered. See details. Defaults to `NULL`, i.e. the full range of each statistic is going to be used.

`col.cov` `Integer vector`. Line color for the covariate trajectory curve of each selected covariate. Defaults to vector of colors of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`lty.cov` `Integer vector`. Line type for the covariate trajectory curve of each selected covariate. Defaults to vector of 1's of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`lwd.cov` `Integer vector`. Line width for the covariate trajectory curve of each selected covariate. Defaults to vector of 1's of length the number of selected covariates. The vector is reused cyclically if it is shorter than the number of selected covariates.

`col` `Integer` scalar. Line color for the trajectory curve of each statistical quantity of interest. Defaults to 1.

`lty` `Integer` scalar. Line type for the trajectory curve of each statistical quantity of interest. Defaults to 1.

`lwd` `Numeric` scalar. Line width for the trajectory curve of each statistical quantity of interest. Defaults to 0.5.

`cex` `Numeric` scalar. Symbol expansion used for titles, captions, and axis labels. Defaults to 0.5.

`add.caption` `Logical` scalar. Should the caption be plotted? Defaults to `FALSE`.

`text.caption` `Character vector` of caption content. Defaults to `NULL`.

`nr` `Integer` scalar of the number of rows in the plot. If `NULL`, defaults to 3.

`nc` `Integer` scalar of the number of columns in the plot. If `NULL`, defaults to 3.

`device` Graphic display device in {`NULL`, "PS", "PDF"}. Defaults to `NULL` (standard output screen). Currently implemented graphic display devices are "PS" (Postscript) or "PDF" (Portable Document Format).

`file` File name for output graphic. Defaults to "Trajectory Plots".

`path` Absolute path (without final (back)slash separator). Defaults to working directory path.

`horizontal` `Logical` scalar. Orientation of the printed image. Defaults to `FALSE`, that is potrait orientation.

`width` `Numeric` scalar. Width of the graphics region in inches. Defaults to 8.5.

`height` `Numeric` scalar. Height of the graphics region in inches. Defaults to 11.

`...` Generic arguments passed to other plotting functions.

Details

The plot shows peeling trajectories of some box descriptive summary statistics and survival output statistics as a function of box support (i.e. peeling steps). It plots peeling trajectories of those only covariates that are used for peeling. It also plots according to the peeling criterion (`peelcriterion`) that is used so that only relevant outputs are plotted. These outputs are: Size (remaining sample size n in the box), Maximum Event-Free Time (MEFT), Minimum Event-Free Probability (MEFP), Log-Hazard Ratio (LHR), Log-Rank Test (LRT), and Concordance Error Rate (CER) if `peelcriterion` in {"lhr", "lrt", "chs"}, or Group Log-Hazard Ratio (GLHR), and Group Concordance Error Rate (GCER) if `peelcriterion` = "grp".

The `range` list includes user-specified ranges of Log-Hazard Ratio (LHR), Log-Rank Test (LRT), and Concordance Error Rate (CER) if `peelcriterion` in {"lhr", "lrt", "chs"}, or Group Log-Hazard Ratio (GLHR), and Group Concordance Error Rate (GCER) if `peelcriterion` = "grp". In the former case, the list should be of the form `range` = `list("lhr"=...,"lrt"=...,"cer"=...)` In the latter case, the list should be of the form `range` = `list("glhr"=...,"gcer"=...)`.

Value

Invisible. None. Displays the plot(s) on the specified `device`.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user plotting function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`predict.sbh` *Predict Function*

Description

S3-method `predict` function to predict the box membership and box vertices on an independent set, using a cross-validated `sbh` fitted object.

Usage

```
  ## S3 method for class 'sbh'
predict(object,
        newdata,
        steps = 1:object$cvfit$cv.nsteps,
        groups = NULL,
        na.action = na.omit, ...)
```

Arguments

`object`	Object of class `sbh` as generated by the main function `sbh`.
`newdata`	A `numeric` `matrix` of same p input covariates as input data `object$X`. If `newdata` is a `numeric` `data.frame`, it is acceptable and will be coerced to a `numeric` `matrix`. Discrete nominal covariates will be treated as ordinal variables. `NA` missing values are not allowed. See details.

`steps`	`Integer vector`. Vector of peeling steps at which to predict the box memberships and box vertices. Defaults to all the peeling steps of sbh object `object`.
`groups`	Same as sbh input: `character` or `numeric vector`, or `factor` of group membership indicator variable, automatically converted into a `factor`. It is of length the sample size, where entries take on group `levels`. Only two groups are allowed at this point. Defaults to `NULL`. See sbh for details.
`na.action`	A function to specify the action to be taken if NAs are found. The default action is na.omit, which leads to rejection of incomplete cases.
`...`	Further generic arguments passed to the predict function.

Details

Only the used covariates of the final sbh object will be retained for the covariates of `newdata`. So, the used covariates of sbh object must be equal or a subset of the the covariates of `newdata`.

Matrix `newdata` is usually an independent dataset drawn from the same population as the input `X` dataset. It can also come from a split input dataset into two sets, usually for validation purposes.

Value

Object of `class` `sbh` (Patient Recursive Survival Peeling), where the `cvfit` list contains the usual values of `sbh` object.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user predict function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).

- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`PRIMsrc.news` *Display the* **PRIMsrc** *Package News*

Description

Function to display the log file `NEWS` of updates of the **PRIMsrc** package.

Usage

```
PRIMsrc.news(...)
```

Arguments

`...`	Further arguments passed to or from other methods.

Value

None.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

---

print.sbh *Print Function*

Description

S3-method `print` function to display the cross-validated estimated values of the `sbh` object.

Usage

```
  ## S3 method for class 'sbh'
print(x, ...)
```

Arguments

`x`	Object of class `sbh` as generated by the main function `sbh`.
`...`	Further generic arguments passed to the print function.

Value

Display of the cross-validated fitted values of its argument.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user print function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.

- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`Real.1` *Real Dataset #1: Clinical Dataset ($p < n$ case)*

Description

Publicly available HIV clinical data from the Women's Interagency HIV cohort Study (WIHS). The entire study enrolled 1164 women. Inclusion criteria of the study are: women at enrolment must be (i) alive, (ii) HIV-1 infected, and (iii) free of clinical AIDS symptoms. Women were followed until the first of the following occurred: (i) treatment initiation (HAART), (ii) AIDS diagnosis, (iii) death, or administrative censoring. The studied outcomes were the competing risks "AIDS/Death (before HAART)" and "Treatment Initiation (HAART)". However, for simplification purposes, only the first of the two competing events (i.e. the time to AIDS/Death), was used. Likewise, for simplification in this clinical dataset example, only complete cases were used. Variables included history of Injection Drug Use ("IDU") at enrollment, African American ethnicity ('Race'), age ('Age'), and baseline CD4 count ('CD4') for a total of $p = 4$ clinical covariates. The question in this dataset example was whether it is possible to achieve a prognostication of patients for AIDS and HAART. See Bacon et al. (2005) and the WIHS website for more details.

Usage

```
data("Real.1", package="PRIMsrc")
```

Format

Dataset consists of a `numeric` `data.frame` containing $n = 485$ complete observations (samples) by rows and $p = 4$ clinical covariates by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

Source

See real data application in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.
- Bacon M.C., von Wyl V., Alden C. et al. (2005). "*The Women's Interagency HIV Study: an observational cohort brings clinical sciences to the bench.*" Clin. Diagn. Lab. Immunol., 12(9):1013-1019.

See Also

Women's Interagency HIV cohort Study website: https://statepi.jhsph.edu/wihs/wordpress/

`Real.2` *Real Dataset #2: Genomic Dataset ($p >> n$ case)*

Description

Publicly available lung cancer genomic data from the Chemores Cohort Study. This data is part of an integrated study of mRNA, miRNA and clinical variables to characterize the molecular distinctions between squamous cell carcinoma (SCC) and adenocarcinoma (AC) in Non Small Cell Lung Cancer (NSCLC) aside large cell lung carcinoma (LCC). Tissue samples were analysed from a cohort of 123 patients, who underwent complete surgical resection at the Institut Mutualiste Montsouris (Paris, France) between 30 January 2002 and 26 June 2006. The studied outcome was the "Disease-Free Survival Time". Patients were followed until the first relapse occurred or administrative censoring. In this genomic dataset, the expression levels of Agilent miRNA probes ($p = 939$) were included from the $n = 123$ cohort samples. The miRNA data contains normalized expression levels. See below the paper by Lazar et al. (2013) and Array Express data repository for complete description of the samples, tissue preparation, Agilent array technology, and data normalization. In addition to the genomic data, five clinical variables, also evaluated on the cohort samples, are included as continuous variable ('Age') and nominal variables ('Type','KRAS.status','EGFR.status','P53.status'). This dataset represents a situation where the number of covariates dominates the number of complete observations, or $p >> n$ case. See Lazar et al. (2013) and the CHEMORES Consortium website for more details.

Usage

```
data("Real.2", package="PRIMsrc")
```

Format

Dataset consists of a `numeric` `data.frame` containing $n = 123$ complete observations (samples) by rows and $p = 939$ genomic covariates by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

Source

See real data application in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.

- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.
- Lazar V. et al. (2013). "*Integrated molecular portrait of non-small cell lung cancers.*" BMC Medical Genomics 6:53-65.

See Also

Array Express data repository at the European Bioinformatics Institute. Accession number: #E-MTAB-1134 (MIR): www.ebi.ac.uk/arrayexpress/

CHEMORES Consortium website: http://www.chemores.ki.se/index.html

---

sbh *Cross-Validated Survival Bump Hunting*

Description

Main end-user function for fitting a Survival Bump Hunting (SBH) model (or Group Survival Bump Hunting (GSBH)). It returns an object of class sbh, as generated by our Patient Recursive Survival Peeling (PRSP) algorithm (or Patient Recursive Group Survival Peeling (PRGSP)), containing cross-validated estimates of the target region (bump) of the input space with end-points statistics of interest. See Dazard and Rao (2014, 2015, 2016, 2021a, 2021b) for details, as well as Dazard et al. (2021c) ans Rao et al. (2020) for applications in Patient Survival Subtyping and Survival Disparity Subtyping.

Usage

```
sbh(X,
    y,
    groups = NULL,
    delta,
    B = 30,
    K = 5,
    A = 1000,
    vs = TRUE,
    vstype = "ppl",
    vsarg = "alpha=1,
             nalpha=1,
             nlambda=100",
    cv = TRUE,
    cvtype = "combined",
    cvarg = "alpha=0.01,
             beta=0.10,
             peelcriterion=\"lrt\",
             cvcriterion=\"cer\"",
    pv = FALSE,
    control = sbh.control(vscons = 0.5,
                          decimals = 2,
                          onese = FALSE,
                          probval = NULL,
                          timeval = NULL,
                          lag = 2,
                          span = 0.10,
                          degree = 2),
    parallel.vs = FALSE,
    parallel.rep = FALSE,
    parallel.pv = FALSE,
    conf = NULL,
    verbose = TRUE,
    seed = NULL)
```

Arguments

X A numeric (n x p) matrix of n observations and p input covariates. If input X is a numeric data.frame, it is acceptable and will be coerced to a numeric matrix. Discrete nominal covariates will be treated as ordinal variables. NA missing values are not allowed.

y n-numeric vector of observed times to event. NA missing values are not allowed.

groups character or numeric vector, or factor of group membership indicator variable, automatically converted into a factor. It is of length the sample size, where entries take on group levels. Only two groups are allowed at this point (see details). Defaults to NULL, i.e. when regular Patient Recursive Survival Peeling (PRSP) is used.

delta n-numeric vector of observed status (censoring) indicator variable.

B Postitive integer of the number of replications of the cross-validation procedure. Defaults to 30.

K Postitive integer of the number of cross-validation folds (partitions) into which the total number of observations (n) should be randomly split. See details.

A Positive integer of the number of permutations used for generating the permutation distribution of the null distribution and the computation of log-rank permutation p-values. Defaults to 1000. Ignored if pv=FALSE.

vs logical scalar. Flag for optional variable (covariate) screening (pre-selection). Defaults to TRUE.

vstype character vector in {"ppl", "pcqr", "spca", "prsp"} of one of the four possible variable screening (pre-selection) procedure. See details below. Defaults to "ppl". Ignored if vs is FALSE.

vsarg Character vector of parameters of the cross-validated variable screening (pre-selection) procedure. Defaults to parameters values of default variable screening (pre-selection) procedure "ppl": vsarg="alpha=1,nalpha=1,nlambda=100". Note that vsarg comes as a character string between double quotes, with comas separated values, and without white spaces. All the following parameters are

ignored if vs is FALSE.
PRSP:

- alpha = numeric data quantile in (0,1) to peel off at each step of the peeling sequence of the PRSP algorithm. Suggests 0.01.
- beta = numeric scalar of minimum box support at the end of the peeling sequence. Suggests 0.10.
- msize = positive integer or NULL to control the model size, i.e the number of screened variables used for fitting the Survival Bump Hunting model. Use a single non-NULL value as the maximum model size (cardinal of subset of top-screened variables) within the allowable range [1,floor(p)]. Alternatively, use msize=NULL to allow the optimal model size to be determined by cross-validation. See below for details. Suggests NULL.
- peelcriterion in {"lhr", "lrt", "chs", "grp"} stands for the peeling criterion used in the rate of increase between in-bump vs out-bump Log-Hazard Ratio (LHR), Log-Rank Test (LRT), or Cumulative Hazard Summary (CHS), and between fixed groups (GRP), respectively (see details). LHR, LRT, and CHS are used in the PRSP algorithm for building a SBH model, while GRP is used in the PRGSP algorithm for building a GSBH model. Suggests "lrt" for SBH and "grp" for GSBH.
- cvcriterion in {"lhr", "lrt", "cer"} stands for the cross-validation criterion Log-Hazard Ratio (LHR), Log-Rank Test (LRT), or Concordance Error Rate (CER), respectively, that is used for optimizing the model size (cardinal of subset of top-screened variables) and the optimal number of peeling steps (optimal peeling sequence length) in the PRSP variable screening procedure. Suggests "cer".

PCQR:

- tau = numeric quantile in [0, 0.5] used in the censored quantile regression model. It is the tuning parameter of the censored quantile loss. It represents the conditional censored quantile of the survival response to be estimated. It includes the absolute loss when tau = 0.5. Suggests 0.5.
- alpha = numeric elasticnet mixing parameter in [0, 1] that controls the relative contribution from the lasso and the ridge penalty. The penalty is defined as (1-alpha)/2||beta||_2^2+alpha||beta||_1. alpha = 1 is the lasso penalty, and alpha = 0 the ridge penalty. If alpha is set to NULL, a vector of values of length nalpha is used, else alpha value is used and nalpha is set to 1. Suggests alpha=1 (lasso).
- nalpha = positive integer of number of alpha values to consider in the grid search. Suggests 1 (see above: lasso).
- nlambda = positive integer of number of elasticnet penalization lambda values to consider in the grid search. Suggests 100.

PPL:

- alpha = numeric elasticnet mixing parameter in [0, 1] that controls the relative contribution from the lasso and the ridge penalty. See R package **glmnet**. The penalty is defined as (1-alpha)/2||beta||_2^2+alpha||beta||_1. alpha = 1 is the lasso penalty, and alpha = 0 the ridge penalty. If alpha is set to NULL, a vector of values of length nalpha is used, else alpha value is used and nalpha is set to 1. Suggests alpha=1 (lasso).
- nalpha = positive integer of number of alpha values to consider in the grid search. Suggests 1 (see above: lasso).

- `nlambda` = positive `integer` of number of elasticnet penalization `lambda` values to consider in the grid search. Suggests 100.

SPCA:

- `n.thres` = number of thresholds to consider in the grid search. It cannot be less than n (sample size). Suggests 20.
- `n.pcs` = number of cross-validation principal components to use in {1,2,3}. It cannot be less than n (sample size) and greater than p (dimensionality of covariates); otherwise, it will be reset to `n.pcs` = p - 1. Suggests 3.
- `n.var` = minimum number of variables to include in determining range for threshold. If cannot be greater than p (dimensionality of covariates); otherwise, it will be reset to `n.var` = p - 1. Suggests 5.

`cv` — `logical` scalar. Flag for optional cross-validation (CV) of variable screening (pre-selection) parameters and Survival Bump Hunting fitting by PRSP algorithm. See below for details. Defaults to `TRUE`.

`cvtype` — `character vector` in {"combined", "averaged"} specifying the cross-validation technique. Defaults to "combined". Ignored if `cv` is `FALSE`.

`cvarg` — `character vector` describing the parameters used in the PRSP algorithm for fitting the Survival Bump Hunting model. Defaults to:
`cvarg="alpha=0.01,beta=0.10,peelcriterion=\"lrt\",cvcriterion=\"cer\""`.
Note that `cvarg` comes as a character string between double quotes, with comas separated values, and without white spaces.

- `alpha` = `numeric` data quantile in (0,1) to peel off at each step of the peeling sequence of the PRSP algorithm. Defaults to 0.01.
- `beta` = `numeric` scalar of minimum box support at the end of the peeling sequence. Defaults to 0.10.
- `peelcriterion` in {"lhr", "lrt", "chs", "grp"} stands for the peeling criterion used in the rate of increase between in-bump vs out-bump Log-Hazard Ratio (LHR), Log-Rank Test (LRT), or Cumulative Hazard Summary (CHS), and between groups (GRP), respectively (see details). LHR, LRT, and CHS are used in the PRSP algorithm for building a SBH model, while GRP is used in the PRGSP algorithm for building a GSBH model. Defaults to "lrt" for SBH and "grp" for GSBH.
- `cvcriterion` in {"lrt", "lhr", "cer"} stands for the cross-validation criterion Log-Hazard Ratio (LHR), Log-Rank Test (LRT), or Concordance Error Rate (CER), respectively, that is used for tuning/optimizing the optimal number of peeling steps. If a SBH model is built (`peelcriterion` in {"lrt", "lhr", "chs"}), all three `cvcriterion` in {"lrt", "lhr", "cer"} are admissible. If a GSBH model is built, only `cvcriterion` in {"cer"} is allowed. Ignored if `cv` is `FALSE`. Defaults to "cer".

if `peelcriterion` is in {"lrt", "lhr", "chs"}, `groups` is automatically set to `NULL`.

`pv` — `logical` scalar. Flag for computation of log-rank permutation p-values. Defaults to `FALSE`.

`control` — Optional function to set ancillary parameters for fitting the Survival Bump Hunting (SBH) model. See `sbh.control` for details.

`parallel.vs` — `logical`. Is parallelization to be performed for variable screening? Defaults to `FALSE`, because it is not implemented yet.

`parallel.rep` — `logical`. Is parallelization to be performed for replications? Defaults to `FALSE`.

`parallel.pv` — `logical`. Is parallelization to be performed for computation of log-rank p-values? Defaults to `FALSE`.

conf — list of 5 fields containing the parameters values needed for creating the parallel backend (cluster configuration). See details below for usage. Optional, defaults to NULL, but all fields are required if used:

- type : character vector specifying the cluster type ("SOCKET", "MPI").
- spec : A specification (character vector or integer scalar) appropriate to the type of cluster.
- homogeneous : logical scalar to be set to FALSE for inhomogeneous clusters.
- verbose : logical scalar to be set to FALSE for quiet mode.
- outfile : character vector of an output log file name to direct the stdout and stderr connection output from the workernodes. "" indicates no redirection.

verbose — logical scalar. Is the output to be verbose? Optional, defaults to TRUE.

seed — Positive integer scalar of the user seed to reproduce all the results. Defaults to NULL.

Details

The main function sbh relies on an optional variable screening (pre-selection) procedure that is run before the actual variable usage (selection) is done at the time of fitting the Survival Bump Hunting (SBH) or Group Survival Bump Hunting (GSBH) model using our PRSP or PRGSP algorithm, respectively. The user can choose between four possible variable screening (pre-selection) procedures (see Dazard and Rao (2021a) for details, as well as Dazard et al. (2021c) for an application in Patient Survival Subtyping):

- Patient Recursive Survival Peeling (PRSP) (by univariate screening of our algorithm)
- Penalized Censored Quantile Regression (PCQR) (by Semismooth Newton Coordinate Descent fiting algorithm adapted from package **hqreg**)
- Penalized Partial Likelihood (PPL) (by Elasticnet Regularization adapted from package **glmnet**)
- Supervised Principal Component Analysis (SPCA) (by Supervised Principal Component adapted from package **superpc**)

NA missing values are not allowed in **PRIMsrc**, because it depends on R package **glmnet**, which doesn't handle missing values. In case of high-dimensional data ($p >> n$), the recommendation is to use PPL or SPCA because of computational efficiency. Variable screening (pre-selection) is done by computing occurrence frequencies of top-ranking variables over the cross-validation folds and replicates. The conservativeness of the procedure is controled by the argument vscons.

The argument K must be bigger than 2 for a regular K-fold cross-validation procedure to work, and should be no less than 3 for a regular procedure to make sense; K $\in$ {5,...,10} is recommended; defaults to K=5. Setting K also specifies the type of cross-validation to be done:

- K = 1 carries no cross-validation out, or set-value when cv = FALSE (see below).
- K $\in$ {2,...,$n-1$} carries out K-fold cross-validation.
- K = n carries out leave-one-out cross-validation.

If cross-validation is done (cv = TRUE, the optimal number of peeling steps (optimal peeling sequence length), and the optimal model size (cardinal of subset of top-screened variables) will be determined by cross-validation. If cv = FALSE, no cross-validation at all will be performed, and the values of K and vscons will both be reset to 1.

The Patient Recursive Group Survival Peeling (PRGSP) algorithm is a derivation of our original Patient Recursive Survival Peeling (PRSP) algorithm to search for (or find an extreme of) outcome *difference* within existing (user-defined) fixed groups of observations. See Dazard and Rao (2021b) for details, as well as Rao et al. (2020) for an application in Survival Disparity Subtyping.

The argument `object$cvarg$peelcriterion` is the peeling criterion that determines what type of bump hunting is done, that is, either using the PRSP or PRGSP algorithm for building a SBH model or GSBH model, respectively. If a regular hunt of *bump difference* is done (SBH model, `peelcriterion` in {"lrt", "lhr", "chs"}), PRSP algorithm is used, and cross-validated bumps are generated between observations from the higher-risk bump (in-bump) versus lower-risk bump (out-bump). If a hunt of (user-specified) fixed *group difference* is done (GSBH model, `peelcriterion` = {"grp"}), PRGSP algorithm is used, and cross-validated bumps are generated between interacting subgroup from the fixed groups and bumps (higher-risk (in-bump) versus lower-risk bump (out-bump)), that is, either between both groups within the higher-risk bump (in-bump), or equivalently, between higher-risk (in-bump) versus lower-risk bump (out-bump) within a given group.

The argument `groups` is to be specified only if a hunt of (user-specified) fixed *group difference* is to be done, i.e. when option `peelcriterion` = {"grp"} and PRGSP are used.

In the PRSP variable screening procedure (`vsarg` of "prsp"), setting option `msize` to a single non-`NULL` value within the allowable range [1,`floor`(p)] will override the cross-validation setting within the variable screening procedure. This could be recommended for high-dimensional data ($p >> n$) to reduce the computational burden. In this situation, we suggest an arbitrary value of `msize` within [1, `floor`(p/5)]. Conversely, setting `msize=NULL` will force the cross-validation within the variable screening procedure by automaticaly generating a vector of model sizes (cardinals of subset of top-screened variables) within the restricted range [1, `floor`(p/5)], which will be used to determine the optimal value of model size.

In fitting the Survival Bump Hunting (SBH) model itself, note that the result contains initial step #0, which corresponds to the entire set of the (training) data. Also, the number of peeling steps that is within the allowable range [1,`ceiling`(`log`(1/n) / `log`(1 - (1/n)))] is further restricted when either of the metaparameter `alpha` or `beta` takes on values other than the smallest possible fraction of the (training) data, i.e. $\frac{1}{n^t}$, where n^t is the training sample size:

- `ceiling`(`log`(`beta`) / `log`(1 - `alpha`)) : `alpha` and `beta` fixed by user
- `ceiling`(`log`(1/n^t) / `log`(1 - `alpha`)) : `alpha` fixed by user and `beta` fixed by data
- `ceiling`(`log`(`beta`) / `log`(1 - (1/n^t))) : `alpha` fixed by data and `beta` fixed by user
- `ceiling`(`log`(1/n^t) / `log`(1 - (1/n^t))) : `alpha` and `beta` fixed by data

When cross-validation is requested (`cv=TRUE`), the function performs a supervised (stratified) random splitting of the observations accounting for the classes/strata provided by `delta` (censoring). This is because it is desireable that the data splitting balances the class distributions of the outcome within the cross-validation splits. For each screening method and for building the final Survival Bump Hunting (SBH) model, all model tuning parameters are simultaneously estimated by cross-validation. The function offers a number of options for the cross-validation to be perfomed: the number of replications `B`; the type of technique; the peeling criterion; and the optimization criterion.

The returned S3-class `sbh` object contains cross-validated estimates of all the decision-rules of used (selected) covariates and all other statistical quantities of interest at each iteration of the peeling sequence (inner loop of the PRSP algorithm). This enables the graphical display of results of profiling curves for model tuning, peeling trajectories, covariate traces and survival distributions (see plotting functions for more details).

In case replicated cross-validations are performed, a "summary report" of the outputs is done over the `B` replicates as follows:

- Even thought the PRSP algorithm uses only one covariate at a time at each peeling step, the reported matrix of "Replicated CV" box decision rules may show more than one covariate being used in a given step depending on the replication. In the end, the reported "Replicated CV" trace values are computed (at each peeling step) as a *single* modal trace value of covariate usage over the `B` replicates. This is also reflected in the "Replicated CV" importance and usage plots of covariate traces.
- Similarly, the reported "Replicated CV" box membership indicators are computed (at each peeling step) as the point-wise modal membership value, that is majority vote, over the `B` replicates (right-hand side of equation #22 in Dazard et al. 2016). The reported "Replicated CV" box support and corresponding box sample size are computed (at each peeling step) based on the above "Replicated CV" box membership indicators (i.e. *not* as equation #23 in Dazard et al. 2016).
- All other reported "Replicated CV" box estimates are computed (at each peeling step) as average statistics over the `B` replicates (i.e. as equation #21 in Dazard et al. 2016), that is, *not* as a single box estimate computed from the "Replicated CV" box membership indicators. This includes the decision rules, the p-values, and all other box statistics. This may result in some apparent discordance if these estimates are re-computed directly from the reported "Replicated CV" box membership indicators.

If the computation of log-rank p-values is desired, then running with the parallelization option is strongly recommended. In case of large ($p > n$) or very large ($p >> n$) datasets, it is also highly recommended to use the parallelization option.

The function `sbh` relies on the R package **parallel** to create a parallel backend within an R session. This enables access to a cluster of compute cores and/or nodes on a local and/or remote machine(s) and scaling-up with the number of CPU cores available and efficient parallel execution. To run a procedure in parallel (with parallel RNG), argument `parallel` is to be set to `TRUE` and argument `conf` is to be specified (i.e. non `NULL`). Argument `conf` uses the options described in function `makeCluster` of the R packages **parallel** and **snow**. **PRIMsrc** supports two types of communication mechanisms between master and worker processes: 'Socket' or 'Message-Passing Interface' ('MPI'). In **PRIMsrc**, parallel 'Socket' clusters use sockets communication mechanisms only (no forking) and are therefore available on all platforms, including Windows, while parallel 'MPI' clusters use high-speed interconnects mechanism in networks of computers (with distributed memory) and are therefore available only in these architectures. A parallel 'MPI' cluster also requires R package **Rmpi** to be installed. Value `type` is used to setup a cluster of type 'Socket' ("SOCKET") or 'MPI' ("MPI"), respectively. Depending on this type, values of `spec` are to be used alternatively:

- For 'Socket' clusters (`conf$type="SOCKET"`), `spec` should be a `character vector` naming the hosts on which to run the job; it can default to a unique local machine, in which case, one may use the unique host name "localhost". Each host name can potentially be repeated to the number of CPU cores available on the local machine. It can also be an `integer` scalar specifying the number of processes to spawn on the local machine; or a list of machine specifications if you have ssh installed (a character value named host specifying the name or address of the host to use).
- For 'MPI' clusters (`conf$type="MPI"`), `spec` should be an `integer` scalar specifying the total number of processes to be spawned across the network of available nodes, counting the workernodes and masternode.

The actual creation of the cluster, its initialization, and closing are all done internally. For more details, see the reference manual of R package **snow** and examples below.

When random number generation is needed, the creation of separate streams of parallel RNG per node is done internally by distributing the stream states to the nodes. For more details, see the vignette of R package **parallel**. The use of a seed allows to reproduce the results within the same type of session: the same seed will reproduce the same results within a non-parallel session or within a parallel session, but it will not necessarily give the exact same results (up to sampling variability) between a non-parallelized and parallelized session due to the difference of management of the seed between the two (see parallel RNG and value of returned seed below).

Value

Object of `class` `sbh` (Patient Recursive Survival Peeling) `list` containing the following 23 fields:

`X` `numeric` `matrix` of original dataset.

`y` `numeric` `vector` of observed failure / survival times.

`groups` `vector` of group membership if algorithm Patient Recursive Group Survival Peeling (PRGSP) is used.

`delta` `numeric` `vector` of observed event indicator in {1,0}.

`B` positive `integer` of the number of replications used in the cross-validation procedure.

`K` positive `integer` of the number of folds used in the cross-validation procedure.

`A` positive `integer` of the number of permutations used for the computation of log-rank permutation p-values.

`vs` `logical` scalar of returned flag of optional variable pre-selection.

`vstype` `character` `vector` of the optional variable pre-selection procdure used.

`vsarg` `list` of parameters used in the pre-selection procedure.

`cv` `logical` scalar of returned flag of optional cross-validation.

`cvtype` `character` `vector` of the cross-validation technique used.

`cvarg` `list` of parameters used in the Survival Bump Hunting procedure.

`pv` `logical` scalar of returned flag of optional computation of log-rank permutation p-values.

`control` `list` of 8 fields of secondary (ancillary) parameters used for fitting the Survival Bump Hunting (SBH) model.

`vscons` `numeric` scalar of conservativeness of the variable screening (pre-selection) procedure.

`onese` `logical` scalar of returned flag of 1-standard error rule.

`decimals` `integer` of the number of user-specified significant decimals.

`probval` `Numeric` scalar of survival probability used.

`timeval` `Numeric` scalar of survival time used.

`cvprofiles` `list` of 10 fields of cross-validated tuning profiles and estimates, each of length `B` (one for each replicate):

- cv.varprofiles: `numeric` `matrix` of cross-validation criterion used for tuning/optimizing the variable screening size in the PRSP variable screening (pre-selection) procedure (`NULL` otherwise). Values are by columns (peeling steps) and replicates (rows).
- cv.varprofiles.mean: `numeric` `vector` of means (across replicates) of the above cross-validation criterion by peeling steps.
- cv.varprofiles.se: `numeric` `vector` of standard errors (across replicates) of the above cross-validation criterion by peeling steps.

- cv.varset.opt: numeric scalar of optimal variable screening size according to the extremum.
- cv.varset.1se: numeric scalar of optimal variable screening size according to 1SE rule.
- cv.stepprofiles: numeric matrix of cross-validation criterion used for tuning/optimizing the peeling sequence length (i.e. number of peeling steps) in the PRSP algorithm. Values are by columns (peeling steps) and replicates (rows).
- cv.stepprofiles.mean: numeric vector of means (across replicates) of the above cross-validation criterion by peeling steps.
- cv.stepprofiles.se: numeric vector of standard errors (across replicates) of the above cross-validation criterion by peeling steps.
- cv.nsteps.opt: numeric scalar of optimal number of peeling steps according to the extremum.
- cv.nsteps.1se: numeric scalar of optimal number of peeling steps according to 1SE rule.

cvfit list with 12 fields of cross-validated SBH output estimates, each of length B (one for each replicate):

- cv.maxsteps: numeric scalar of maximal number of peeling steps (counting step #0 - see Details section).
- cv.nsteps: numeric scalar of optimal number of peeling steps (counting step #0 - see Details section).
- cv.boxind: logical matrix in TRUE, FALSE of individual observation box membership indicator (columns) for all peeling steps (rows).
- cv.boxind.size: numeric vector of box sample size for all peeling steps.
- cv.boxind.support: numeric vector of box support for all peeling steps.
- cv.rules: data.frame of decision rules on the covariates (columns) for all peeling steps (rows).
- cv.screened: numeric vector of screened (pre-selected) covariates, indexed in reference to original index.
- cv.trace: numeric vector of the modal trace values of covariate usage for all peeling steps.
- cv.sign: numeric vector in {-1,+1} of directions of peeling for all used (selected) covariates.
- cv.used: numeric vector of covariates used (selected) for peeling, indexed in reference to original index.
- cv.stats: numeric matrix of box endpoint descriptive statistics of interest (columns) at all peeling steps (rows). The output varies according to the peeling criterion (peelcriterion) used (only relevant outputs are returned).
- cv.pval: list with 2 fields of two vectors. The first cvfit$pval is a numeric vector for log-rank permutation p-values. The second cvfit$seed is is an integer scalar of the seed if parallelization is used, or an integer vector of A values of seeds, one for each permutation, if parallelization is not used.

success logical scalar of the returned flag of success at fitting the SBH model.

seed User seed. An integer scalar if parallelization is used, or an integer vector of B values, one for each replication, if parallelization is not used.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

Main end-user function for fitting the Survival Bump Hunting model.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

See Also

- `sbh.control`
- `makeCluster` (R package **parallel**)
- `glmnet`, `cv.glmnet` (R package **glmnet**)

- hqreg, cv.hqreg (R package **hqreg**)
- superpc.cv (R package **superpc**)

Examples

```
#===================================================
# Loading the library and its dependencies
#===================================================
library("PRIMsrc")

## Not run: 
    #===================================================
    # PRIMsrc Package news
    #===================================================
    PRIMsrc.news()

    #===================================================
    # PRIMsrc Package citation
    #===================================================
    citation("PRIMsrc")

    #===================================================
    # Demo with a synthetic dataset
    # Use help for descriptions
    #===================================================
    data("Synthetic.1", package="PRIMsrc")
    ?Synthetic.1

## End(Not run)

#===================================================
# Simulated dataset #1 (n=250, p=3)
# Peeling criterion = LRT
# Cross-Validation criterion = LRT
# With Combined Cross-Validation (RCCV)
# Without Replications (B = 1)
# Without variable screening (pre-selection)
# Without computation of log-rank \eqn{p}-values
# Without parallelization
#===================================================
data("Synthetic.1", package="PRIMsrc")
synt1 <- sbh(X = Synthetic.1[ , -c(1,2), drop=FALSE],
             y = Synthetic.1[ ,1, drop=TRUE],
             groups = NULL,
             delta = Synthetic.1[ ,2, drop=TRUE],
             B = 1,
             K = 3,
             vs = FALSE,
             cv = TRUE,
             cvtype = "combined",
             cvarg = "alpha=0.05,
                      beta=0.10,
                      peelcriterion=\"lrt\",
                      cvcriterion=\"lrt\"",
             pv = FALSE,
             control = sbh.control(probval = 0.5),
             parallel.vs = FALSE,
             parallel.rep = FALSE,
             parallel.pv = FALSE,
             conf = NULL,
             verbose = TRUE,
             seed = 123)

summary(object = synt1)
print(x = synt1)

n <- 100
p <- length(synt1$cvfit$cv.used)
x <- matrix(data = runif(n = n*p, min = 0, max = 1),
            nrow = n, ncol = p, byrow = FALSE,
            dimnames = list(1:n, paste("X", 1:p, sep="")))
synt1.pred <- predict(object = synt1,
                      newdata = x,
                      steps = synt1$cvfit$cv.nsteps)

plot(x = synt1,
     main = paste("Scatter plot for model #1", sep=""),
     proj = synt1$cvfit$cv.used[c(1,2)],
     steps = synt1$cvfit$cv.nsteps,
     pch = 16, cex = 0.5, col = c(1,2),
     boxes = TRUE,
     col.box = 2, lty.box = 2, lwd.box = 1,
     add.caption.box = TRUE,
     text.caption.box = paste("Step: ", synt1$cvfit$cv.nsteps, sep=""),
     device = NULL)

plot_profile(object = synt1,
             main = "Cross-validated tuning profiles for model #1",
             pch = 20, col = 1, lty = 1, lwd = 0.5, cex = 0.5,
             add.sd = TRUE,
             add.profiles = TRUE,
             add.caption = TRUE,
             text.caption = c("Mean","Std. Error"),
             device = NULL)

plot_traj(object = synt1,
          main = paste("Cross-validated peeling trajectories for model #1", sep=""),
          col = 1, lty = 1, lwd = 0.5, cex = 0.5,
          toplot = synt1$cvfit$cv.used,
          device = NULL)

plot_trace(object = synt1,
           main = paste("Cross-validated trace plots for model #1", sep=""),
           xlab = "Box Support", ylab = "Covariate Range (centered)",
           col = 1, lty = 1, lwd = 0.5, cex = 0.5,
           toplot = synt1$cvfit$cv.used,
           center = TRUE, scale = FALSE,
           device = NULL)

plot_km(object = synt1,
        main = paste("Cross-validated probability curves for model #1", sep=""),
        xlab = "Time", ylab = "Probability",
        ci = TRUE,
        col = c(1,2), lty = 1, lwd = 0.5, cex = 0.5,
        steps = 1:synt1$cvfit$cv.nsteps,
        plot.type = "bumps",
        bump.reference = "in-bump",
        group.reference = levels(synt1$groups)[1],
        add.caption = TRUE,
        text.caption = c("out-bump","in-bump"),
        device = NULL)

## Not run:
    #===================================================
    # Examples of parallel backend parametrization
    #===================================================
    if (require("parallel")) {
       cat("'parallel' is attached correctly \n")
    } else {
       stop("'parallel' must be attached first \n")
    }
    #===================================================
    # Ex. #1 - Multicore PC
    # Running WINDOWS
    # SOCKET communication cluster
    # Shared memory parallelization
    #===================================================
    cpus <- parallel::detectCores(logical = TRUE)
    conf <- list("spec" = rep("localhost", cpus),
                 "type" = "SOCKET",
                 "homo" = TRUE,
                 "verbose" = TRUE,
                 "outfile" = "")
    #===================================================
    # Ex. #2 - Master node + 3 Worker nodes cluster
    # All nodes equipped with identical setups of multicores
    # (8 core CPUs per machine for a total of 32)
    # SOCKET communication cluster
    # Distributed memory parallelization
    #===================================================
    masterhost <- Sys.getenv("HOSTNAME")
    slavehosts <- c("compute-0-0", "compute-0-1", "compute-0-2")
    nodes <- length(slavehosts) + 1
    cpus <- 8
    conf <- list("spec" = c(rep(masterhost, cpus),
                            rep(slavehosts, cpus)),
                 "type" = "SOCKET",
                 "homo" = TRUE,
                 "verbose" = TRUE,
                 "outfile" = "")
    #===================================================
    # Ex. #3 - Enterprise Multinode Cluster w/ multicore/node
    # Running LINUX with SLURM scheduler
    # MPI communication cluster
    # Distributed memory parallelization
    # Below, variable 'cpus' is the total number of requested
    # tasks (threads/CPUs), which is specified from within a
    # SLURM script.
    #==================================================
    if (require("Rmpi")) {
        print("Rmpi is loaded correctly \n")
```

```
} else {
    stop("Rmpi must be installed first to use MPI\n")
}
cpus <- as.numeric(Sys.getenv("SLURM_NTASKS"))
conf <- list("spec" = cpus,
             "type" = "MPI",
             "homo" = TRUE,
             "verbose" = TRUE,
             "outfile" = "")
#===================================================
# Simulated dataset #1 (n=250, p=3)
# Peeling criterion = LRT
# Cross-Validation criterion = LRT
# With Combined Cross-Validation (RCCV)
# With Replications (B = 30)
# With PPL variable screening (pre-selection)
# With computation of log-rank \eqn{p}-values
# With parallelization
#===================================================
data("Synthetic.1", package="PRIMsrc")
synt1 <- sbh(X = Synthetic.1[ , -c(1,2), drop=FALSE],
             y = Synthetic.1[ ,1, drop=TRUE],
             groups = NULL,
             delta = Synthetic.1[ ,2, drop=TRUE],
             B = 30,
             K = 5,
             A = 1000,
             vs = TRUE,
             vstype = "ppl",
             vsarg = "alpha=1,
                      nalpha=1,
                      nlambda=100",
             cv = TRUE,
             cvtype = "combined",
             cvarg = "alpha=0.01,
                      beta=0.10,
                      peelcriterion=\"lrt\",
                      cvcriterion=\"lrt\"",
             pv = TRUE,
             control = sbh.control(probval = 0.5,
                                   vscons = 0.5),
             parallel.vs = FALSE,
             parallel.rep = TRUE,
             parallel.pv = TRUE,
             conf = conf,
             verbose = TRUE,
             seed = 123)
#===================================================
# Simulated dataset #4 (n=100, p=1000)
# Peeling criterion = LRT
# Cross-Validation criterion = CER
# With Combined Cross-Validation (RCCV)
# With Replications (B = 30)
# With PRSP variable screening (pre-selection)
# With computation of log-rank \eqn{p}-values
# With parallelization
#===================================================
    data("Synthetic.4", package="PRIMsrc")
    synt4 <- sbh(X = Synthetic.4[ , -c(1,2), drop=FALSE],
                 y = Synthetic.4[ ,1, drop=TRUE],
                 groups = NULL,
                 delta = Synthetic.4[ ,2, drop=TRUE],
                 B = 30,
                 K = 5,
                 A = 1000,
                 vs = TRUE,
                 vstype = "prsp",
                 vsarg = "alpha=0.01,
                          beta=0.10,
                          msize=NULL,
                          peelcriterion=\"lrt\",
                          cvcriterion=\"cer\"",
                 cv = TRUE,
                 cvtype = "combined",
                 cvarg = "alpha=0.01,
                          beta=0.10,
                          peelcriterion=\"lrt\",
                          cvcriterion=\"cer\"",
                 pv = TRUE,
                 control = sbh.control(probval = 0.5,
                                       vscons = 0.5),
                 parallel.vs = FALSE,
                 parallel.rep = TRUE,
                 parallel.pv = TRUE,
                 conf = conf,
                 verbose = TRUE,
                 seed = 123)

## End(Not run)
```

`sbh.control` *Parameters Control Function*

Description

End-user function to set ancillary parameters of main end-user function sbh for fitting a Survival Bump Hunting (SBH) model. It is used to set some variable screening parameters, optional formats and outputs of sbh, as well as internally to tune the scatterplot smoother used for finding cross-validated model selection/tuning profile extremum.

Usage

```
sbh.control(vscons = 0.5,
            decimals = 2,
            onese = FALSE,
            probval = NULL,
            timeval = NULL,
            lag = 2,
            span = 0.10,
            degree = 2)
```

Arguments

`vscons` — `numeric` scalar in [1/`K`, 1], specifying the conservativeness of the variable screening (pre-selection) procedure, where 1/`K` is the least conservative and 1 is the most. Defaults to 0.5.

`decimals` — Positive `integer` of the number of user-specified significant decimals to output results. Defaults to 2.

`onese` — `logical` scalar. Flag for using the 1-standard error rule instead of extremum value of the cross-validation criterion when tuning/optimizing model parameters. Defaults to `FALSE`.

`probval` — `numeric` scalar in [0, 1] of the survival probability at which we want to get the endpoint box survival time. Defaults to `NULL` (i.e. maximal survival probability value is used).

`timeval` — `numeric` scalar of the survival time at which we want to get the endpoint box survival probability. Defaults to `NULL` (i.e. maximal survival time value is used).

`lag` — Positive `integer` indicating which lag to use in the lagged and iterated difference function. Defaults to 2.

`span` — `numeric` scalar in [0, 1], specifying the degree of smoothing in the internal `stats::loess` function. Defaults to 0.10. If `span` is too small with respect to the number of peeling steps (Adjusted maximum peeling length), choose a larger value such that `floor`(number of peeling steps * `span`) > 0.

`degree` — Positive `integer` indicating the degree of the polynomials (normally 1 or 2) to be used in the internal `stats::loess` function. Here, degree 0 is not also allowed unlike in `stats::loess`). Defaults to 2.

Details

Example of `vscons` values for pre-selection are as follows:

- ‘1.0‘ represents a presence in all the folds (unanimity vote)
- ‘0.5‘ represents a presence in at least half of the folds (majority vote)
- ‘1/`K`‘ represents a presence in at least one of the folds (minority vote)

Although any value in the interval [1/`K`,1] is accepted, we recommand using the interval [1/`K`, 1/2] to avoid excessive conservativeness. Final variable usage (selection) is done at the time of fitting the Survival Bump Hunting (SBH) model itself using our PRSP algorithm on previously screened variables by collecting those variables that have the maximum occurrence frequency in each peeling step over cross-validation folds and replicates.

Value

A list of 8 components.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user function to be used with `sbh`.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

See Also

- `sbh`
- `diff` (R package **base**)
- `loess` (R package **stats**)

summary.sbh *Summary Function*

Description

S3-method `summary` function to summarize the main parameters used to generate the `sbh` object.

Usage

```
  ## S3 method for class 'sbh'
summary(object, ...)
```

Arguments

`object`	Object of class `sbh` as generated by the main function `sbh`.
`...`	Further generic arguments passed to the summary function.

Value

Summarizes the main parameters used to generate its argument.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Note

End-user summary function.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).

- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`Synthetic.1` *Synthetic Dataset #1: $p < n$ case*

Description

Dataset from simulated regression survival model #1 as described in Dazard et al. (2015). Here, the regression function uses all of the predictors, which are also part of the design matrix. Survival time was generated from an exponential model with rate parameter λ (and mean 1/λ) according to a Cox-PH model with hazard exp(eta), where eta(.) is the regression function. Censoring indicator were generated from a uniform distribution on [0, 3]. In this synthetic example, all covariates are continuous, i.i.d. from a multivariate uniform distribution on [0, 1].

Usage

```
data("Synthetic.1", package="PRIMsrc")
```

Format

Each dataset consists of a `numeric matrix` containing $n = 250$ observations (samples) by rows and $p = 3$ variables by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

Source

See simulated survival model #1 in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`Synthetic.1b` *Synthetic Dataset #1b: $p < n$ case*

Description

Dataset from simulated regression survival model #1b as described in Dazard et al. (2015). Here, the regression function uses all of the predictors, which are also part of the design matrix. In this example, the signal is limited to a box-shaped region R of the predictor space. Survival time was generated from an exponential model with rate parameter λ (and mean 1/λ) according to a Cox-PH model with hazard exp(eta), where eta(.) is the regression function. Censoring indicator were generated from a uniform distribution on [0, 3]. In this synthetic example, all covariates are continuous, i.i.d. from a multivariate uniform distribution on [0, 1].

Usage

```
data("Synthetic.1b", package="PRIMsrc")
```

Format

Each dataset consists of a `numeric` `matrix` containing $n = 250$ observations (samples) by rows and $p = 3$ variables by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`
- "Michael Choe, M.D." `<mjc206@case.edu>`
- "Michael LeBlanc, Ph.D." `<mleblanc@fhcrc.org>`
- "Alberto Santana, MBA." `<ahs4@case.edu>`
- "J. Sunil Rao, Ph.D." `<Rao@biostat.med.miami.edu>`

Maintainer: "Jean-Eudes Dazard, Ph.D." `<jean-eudes.dazard@case.edu>`

Source

See simulated survival model #1b in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.

- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`Synthetic.2` *Synthetic Dataset #2: $p < n$ case*

Description

Dataset from simulated regression survival model #2 as described in Dazard et al. (2015). Here, the regression function uses some informative predictors. The rest represent un-informative noisy covariates, which are not part of the design matrix. Survival time was generated from an exponential model with rate parameter λ (and mean $1/\lambda$) according to a Cox-PH model with hazard exp(eta), where eta(.) is the regression function. Censoring indicator were generated from a uniform distribution on [0, 3]. In this synthetic example, all covariates are continuous, i.i.d. from a multivariate uniform distribution on [0, 1].

Usage

```
data("Synthetic.2", package="PRIMsrc")
```

Format

Each dataset consists of a `numeric` `matrix` containing $n = 250$ observations (samples) by rows and $p = 3$ variables by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <`jean-eudes.dazard@case.edu`>
- "Michael Choe, M.D." <`mjc206@case.edu`>
- "Michael LeBlanc, Ph.D." <`mleblanc@fhcrc.org`>
- "Alberto Santana, MBA." <`ahs4@case.edu`>
- "J. Sunil Rao, Ph.D." <`Rao@biostat.med.miami.edu`>

Maintainer: "Jean-Eudes Dazard, Ph.D." <`jean-eudes.dazard@case.edu`>

Source

See simulated survival model #2 in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2021d). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" (in prep).
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

`Synthetic.3` *Synthetic Dataset #3: $p < n$ case*

Description

Dataset from simulated regression survival model #3 as described in Dazard et al. (2015). Here, the regression function does not include any of the predictors. This means that none of the covariates is informative (noisy), and are not part of the design matrix. Survival time was generated from an exponential model with rate parameter λ (and mean $1/\lambda$) according to a Cox-PH model with hazard exp(eta), where eta(.) is the regression function. Censoring indicator were generated from a uniform distribution on [0, 3]. In this synthetic example, all covariates are continuous, i.i.d. from a multivariate uniform distribution on [0, 1].

Usage

```
data("Synthetic.3", package="PRIMsrc")
```

Format

Each dataset consists of a `numeric` `matrix` containing $n = 250$ observations (samples) by rows and $p = 3$ variables by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

Source

See simulated survival model #3 in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).
- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

Synthetic.4 *Synthetic Dataset #4: $p > n$ case*

Description

Dataset from simulated regression survival model #4 as described in Dazard et al. (2015). Here, the regression function uses 1/10 of informative predictors in a $p > n$ situation with $p = 1000$ and $n = 100$. The rest represents non-informative noisy covariates, which are not part of the design matrix. Survival time was generated from an exponential model with rate parameter λ (and mean 1/λ) according to a Cox-PH model with hazard exp(eta), where eta(.) is the regression function. Censoring indicator were generated from a uniform distribution on [0, 2]. In this synthetic example, all covariates are continuous, i.i.d. from a multivariate standard normal distribution.

Usage

```
data("Synthetic.4", package="PRIMsrc")
```

Format

Each dataset consists of a `numeric` `matrix` containing $n = 100$ observations (samples) by rows and $p = 1000$ variables by columns, not including the censoring indicator and (censored) time-to-event variables. It comes as a compressed Rda data file.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. This project was partially funded by the National Institutes of Health NIH - National Cancer Institute (R01-CA160593) to J-E. Dazard and J.S. Rao.

Author(s)

- "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>
- "Michael Choe, M.D." <mjc206@case.edu>
- "Michael LeBlanc, Ph.D." <mleblanc@fhcrc.org>
- "Alberto Santana, MBA." <ahs4@case.edu>
- "J. Sunil Rao, Ph.D." <Rao@biostat.med.miami.edu>

Maintainer: "Jean-Eudes Dazard, Ph.D." <jean-eudes.dazard@case.edu>

Source

See simulated survival model #4 in Dazard et al., 2015.

References

- Dazard J-E. and Rao J.S. (2021a). "*Variable Selection Strategies for High-Dimensional Recursive Peeling-Based Survival Bump Hunting Models.*" (in prep).
- Dazard J-E. and Rao J.S. (2021b). "*Group Bump Hunting by Recursive Peeling-Based Methods: Application to Survival/Risk Predictive Models.*" (in prep).

- Dazard J-E., Choe M., Pawitan Y., and Rao J.S. (2021c). "*Identification and Characterization of Informative Prognostic Subgroups by Survival Bump Hunting.*" (in prep).
- Rao J.S., Huilin Y., and Dazard J-E. (2020). "*Disparity Subtyping: Bringing Precision Medicine Closer to Disparity Science.*" Cancer Epidemiology Biomarkers & Prevention, 29(6 Suppl):C018.
- Yi C. and Huang J. (2017). "*Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression.*" J. Comp Graph. Statistics, 26(3):547-557.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2016). "*Cross-validation and Peeling Strategies for Survival Bump Hunting using Recursive Peeling Methods.*" Statistical Analysis and Data Mining, 9(1):12-42.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2015). "*R package PRIMsrc: Bump Hunting by Patient Rule Induction Method for Survival, Regression and Classification.*" In JSM Proceedings, Statistical Programmers and Analysts Section. Seattle, WA, USA. American Statistical Association IMS - JSM, p. 650-664.
- Dazard J-E., Choe M., LeBlanc M., and Rao J.S. (2014). "*Cross-Validation of Survival Bump Hunting by Recursive Peeling Methods.*" In JSM Proceedings, Survival Methods for Risk Estimation/Prediction Section. Boston, MA, USA. American Statistical Association IMS - JSM, p. 3366-3380.

Index